**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

*IJCSMC, Vol. 3, Issue. 4, April 2014, pg.503 – 508*

**RESEARCH ARTICLE**

# Improvement of Expectation Maximization Clustering using Select Attribute

**Rupali Bhondave[1], Madhura Kalbhor[1], Supriya Shinde[1], K. Rajeswari[2]**

[1]ME Student, Pimpri Chinchwad College of Engineering, Nigdi, Pune, India

[2]Associate Prof., PCCOE Pune & Ph.D Research Scholar, SASTRA University, Tanjore, India

[1] hardikad38@gmail.com ; [2] raji.pccoe@gmail.com

***Abstract:- Data mining is the process of extracting valuable information from various sources of data and produces knowledge. For mining data, WEKA tool is used. In WEKA there are various processes to produce knowledge, like Preprocess, Classification, Clustering, Select Attribute, and Association etc.***

***This paper focuses clustering Technique. Clustering is a technique by which we can categorize similar objects or dissimilar object. There are various algorithms in clustering. A method attribute selection for experimentation on Expectation Maximization (EM) clustering is used. In attribute selection we used Best First Search (BFS), Random Search for EM clustering, which gives better results than the result obtain without using attribute selection method.***

***Keywords:- Data Mining; WEKA; Clustering; EM; BFS; Random search***

## I. INTRODUCTION

Data Mining [10] is the data analysis technique where the data is searched in automated way to solve the problems. It discovers patterns from large dataset. There are various techniques in data mining like regression, clustering, classification. In this paper clustering is used because it is applicable in various real time applications such as banking domain etc.

 

## II. LITERATURE REVIEW

*2.1 Clustering [9]*

Clustering is grouping of objects to find out whether there is some relationship existing between the objects. It is a task through which data should be explored and used for statistical data analysis; this data can be used in various applications like machine learning, pattern recognition, and information retrieval. Different clustering methods generate various types of clusters on same dataset. The partitioning is not performed by human, but by clustering algorithms.

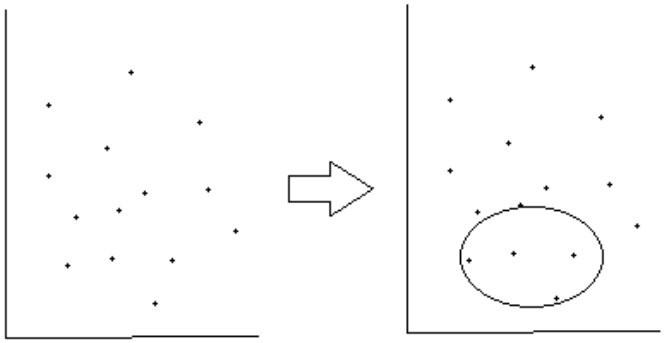

Fig1.2:-Formation of clusters.

Figure1.2 above shows clustering the similar objects,two or more objects belong to same cluster if they are close to according to some distance, this is called as Distance Based Clustering , an another kind of clustering is Conceptual Clustering, where two or more objects are grouped according to a comman concept, not according to distance.There are various clustering algoritms in Weka like Cobweb, DBScan, EM, SimpleKMeans etc.

*2.3 Attibute Selection[8]*

It is a process of selecting subset of relevant features for model creation This is also called as feature selection or variable selection.Feature selection is also usefull in data analysis , It show which features are important and relationship between features for further analysis. Feature selection is particularly important for data sets with large numbers of features e.g. classification problems in molecular biology may involve thousands of features .In supervised learning , feature selection improves the performance of classifier in given dataset but in unsupervised learning, feature selection has very little attention . Our aim is to select best attributes to improve the time of clustering. In WEKA for attribute selection panel to be specified attribute subset evaluator and search method.

In attribute subset evaluator, It takes the subset of attributes and returns numerical measure that guide search. *CfsSubsetEval* predicts each attribute individually and degree of redundancy among attribute, select the attributes which are highly correlated with the class. Search method travers the attribute space to find a good subset. There are various search methods like BestFirst,RandomSearch, and Ranker etc.

*2.4 EM algorithm[9]*

EM algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models[8].EM starts with initial parameters and perform clustering until clustering cannot be improve.

The EM iteration consist of two steps expectation (E) step and maximization (M) steps.
**The Expectation (E) Step: -** Each object assign to clusters with the center that is closest to the object. Assignment of object should be belonging to closest cluster.

 

**The Maximization (M) step: -**For given cluster assignment, for each cluster algorithm adjust the center so that, the sum of the distance from object and new center is minimized.

The result of the cluster analysis is to class indices. The value in class indices, where a value '0' refers to the first cluster and '1'refer to the second cluster etc.

Advantages

1. Gives extremely useful result for real word application such as banking, medical etc.
2. EM handles many learning problem in data mining.
3. It converge to local maximum

Disadvantages

1. EM algorithm can be very costly if the number of distribution is large.

## III. RESOURCES USED

*3.1 WEKA [3]*

It stands for Environment for Knowledge Learning It was developed by the University of Waikato. It support data mining tasks such as data pre-processing, clustering, classification, regression and feature selection. The workflow of WEKA:-

**Input- Pre-processing - Data Mining- Knowledge**

Fig 3.1 Below shows the Graphical User Interface (GUI) of WEKA.


*3.2Dataset:-*

For analysis three data sets are Diabeties disease having 9 attribute and 768 instances , Heart disease having 14 attribute and 123 instances , Visa details [7] having 24 attribute and 3400 instances.

 

## IV. PROPOSED WORK

The figure shows the flow of experiment performed. In this work a combination of clustering techniques with Select Attribute is verified.

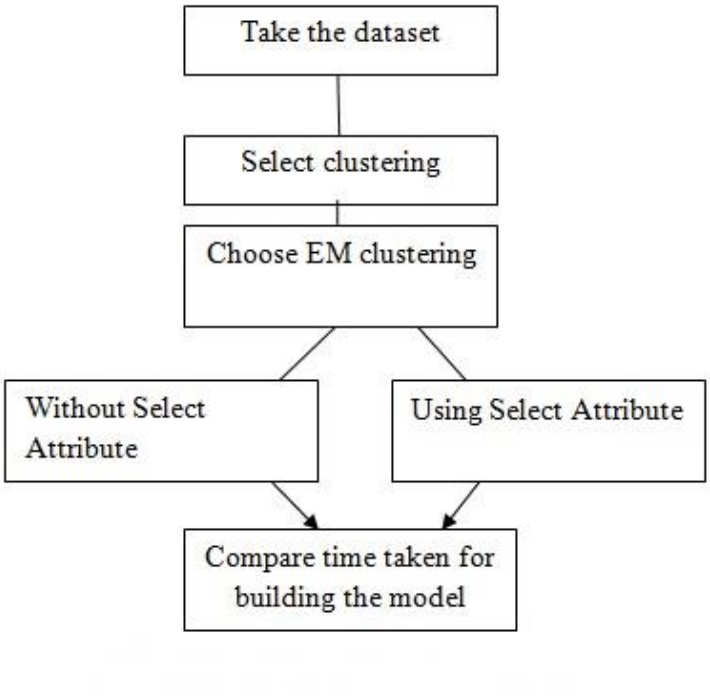


Fig 4:- workflow of clustering process

## V. RESULT AND DISCUSSION

In below tables, we perform some data mining operation on these dataset like clustering, using EM algorithm. First approach is use EM clustering without attribute selection on these dataset. Second approach is select attribute and then performs EM clustering on selected attribute.

| Dataset | Time Taken for EM clustering |
|---|---|
| Diabetes | 7.36sec |
| Heart | 1.03sec |
| Visa Details | 80.52sec |

Table 5.1:-Time Taken for EM clustering by different dataset

 

| Dataset | Time Taken for EM clustering(using Select Attribute) |
|---|---|
| Diabetes | 0.7sec |
| Heart | 0.36sec |
| Visa Details | 1.51sec |

Table 5.2:-Time Taken for EM clustering (using Select Attribute) by different dataset

It is observed that when attributes are select using the BFS and Random Search method, time taken while building model is less.

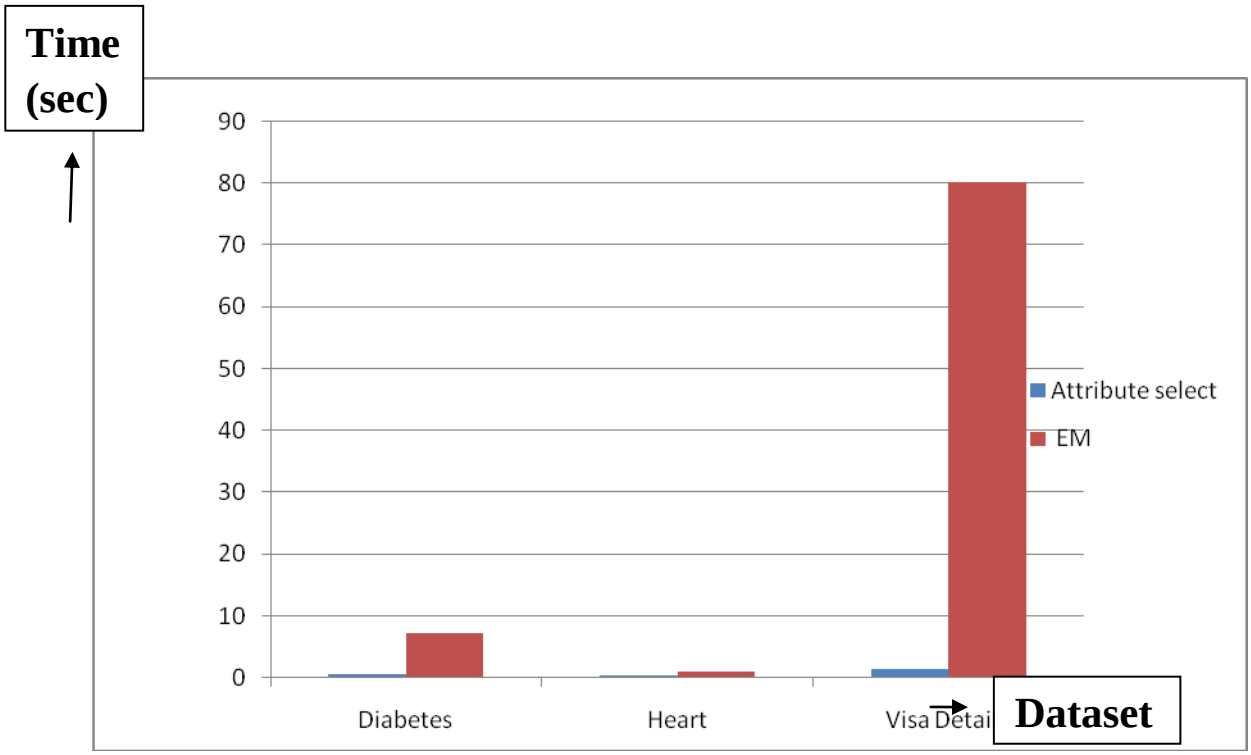


Figure 5.1:-Time taken for EM Clustering by different dataset (with or without using attribute selection)

In above figure 5.1 there are three different dataset are Pima Indian Diabetes from Diabetes, Switzerland from Heart, and Visa Details [7]. EM clustering algorithm takes time respectively as 7.36sec, 1.03sec, 80.52sec, where as if we select attribute selection with search method as BFS or Random Search, EM model takes time respectively as 0.7sec, 0.36sec, 1.51sec.

## VI. CONCLUSION AND FUTURE WORK

In this paper we combine the EM clustering with select attribute method on three dataset like Pima Indian Diabetes, Switzerland Heart Disease and Visa Details etc. It is found that performance of clustering is varies with different datasets. According to results factor affects is time required while building model is less.

Our future work will focus on combine select attribute with classification which improves the efficiency of classification.

 

## REFERENCES

[1] Narendra Sharma , Aman Bajpai , Mr. Ratnesh Litoriya"*Comparison the various clustering algorithms of weka tools*" Department of computer science, Jaypee University Engg. & Technology.

[2] Martin H.C. Law, IEEE, Mario A.T. Figueiredo, Senior Member, IEEE, and Anil K. Jain, Fellow, IEEE- "*Simultaneous Feature Selection and Clustering Using Mixture Models*" Pattern Analysis,vol 26,september 2004.

[3]Dr. V. Vaithiyanathan , K.Rajeswari, Rahul Pitale, Kapil Tajane "*Performance Improvement using Integration of Association Rule Mining and Classification Techniques*" International Journal of Scientific & Engineering Research, Vol 4, May-2013

[4]Yuni Xia, Bowei Xi –"*Conceptual Clustering Categorical Data with Uncertainty* "Indiana University Purdue University Indianapolis, IN 46202, USA

[5] Sanjoy Dasgupta "*Performance guarantees for hierarchical clustering*" Department of Computer Science and Engineering University of California.

[6] A. P. Dempster; N. M. Laird; D. B. Rubin "*Maximum Likelihood from Incomplete Data via the EM Algorithm*" Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38.

[7] www.http://data.gov.in

[8] Ian H.Witten and Elbe Frank, (2005) "*Datamining Practical Machine Learning Tools and Techniques,*" Second Edition, Morgan Kaufmann, San Fransisco.

[9] J. Han and M. Kamber, (2000) "*Data Mining: Concepts and Techniques*," Morgan Kaufmann.

[10] Weka: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka.

 